ST405

Canonical Correlation Analysis

for Steel Industry Energy Consumption Data Set

Maheshika Maduwanthi
S18841

Introduction

The dataset used in this study comes from DAEWOO Steel Co. Ltd in Gwangyang, South Korea, a leading producer of coils, steel plates, and iron plates. It includes detailed information on the company's electricity consumption, stored in a cloud-based system, along with energy consumption data from the Korea Electric Power Corporation (pccs.kepco.go.kr), providing daily, monthly, and annual usage statistics. The dataset covers various operational and environmental factors, allowing for a thorough analysis of energy efficiency in steel production.

The main question of this study is to explore how energy consumption is related to various operational and environmental factors in the steel industry. Specifically, the aim is to understand which factors influence energy usage and how these relationships can help optimize energy efficiency, reduce costs, and minimize environmental impacts.

This study aims to apply Canonical Correlation Analysis (CCA) to the energy consumption data from DAEWOO Steel Co. Ltd. The analysis will explore the correlations between two sets of variables: one set related to energy usage and reactive power, and another set related to environmental factors and power factors. The goal is to uncover the underlying relationships that influence energy efficiency within the steel production process.

Hypotheses

The study is guided by the following hypotheses:

- There are significant correlations between the variables related to energy usage and reactive power (Set 1) and the variables related to environmental factors and power factors (Set 2).
- Specific combinations of variables in Set 1 are strongly correlated with specific combinations of variables in Set 2.
- CCA results will reveal insights to optimize energy consumption in the steel industry.

This study is important for several reasons:

- **Economic Impact**: Understanding relationships between energy consumption and operational factors can identify cost-saving opportunities.
- **Environmental Impact**: Reducing energy consumption helps lower greenhouse gas emissions, supporting sustainability goals.
- **Academic Contribution**: This study applies Canonical Correlation Analysis to a real-world industrial dataset, demonstrating its utility in identifying complex multivariate relationships in industrial settings.

Methodology

The dataset contains 35,040 instances and 9 features, with a mix of real and categorical data. It is collected from a smart small-scale steel industry in South Korea. The data includes various operational and environmental factors recorded daily, monthly, and annually.

Variable Name	Role	Type	Description	Units	Missing Values
date	Other	Date			no
Usage_kWh	Feature	Continuous	Industry Energy Consumption	kWh	no
Lagging_Current_Reactive.Power_kVarh	Feature	Continuous		kVarh	no
Leading_Current_Reactive_Power_kVarh	Feature	Continuous		kVarh	no
CO2(tCO2)	Feature	Continuous		ppm	no
Lagging_Current_Power_Factor	Feature	Continuous		%	no
Leading_Current_Power_Factor	Feature	Continuous		%	no
NSM	Feature	Integer		s	no
WeekStatus	Feature	Categorical	Weekend (0) or a Weekday(1)		no
Day_of_week	Feature	Categorical	Sunday, Monday, ..., Saturday		no
Load_Type	Target	Categorical	Light Load, Medium Load, Maximum Load		no

Statistical Methods Used

- **Canonical Correlation Analysis (CCA)**: This multivariate statistical method is used to explore the relationships between two sets of variables. In this study, CCA is employed to identify the correlations between a set of variables related to energy usage and reactive power (Set 1) and a set of variables related to environmental factors and power factors (Set 2).
- **Standardization**: Before performing CCA, the data is standardized to ensure that each variable contributes equally to the analysis.
- **Wilks' Lambda, Hotelling-Lawley Trace, Pillai's Trace, and Roy's Greatest Root Tests**: These tests are used to assess the significance of the canonical correlations identified by the CCA.

Results and Discussion

The Canonical Correlation Analysis (CCA) was conducted to explore the relationships between two sets of variables:

- **Set 1 (X)**: Related to energy usage and reactive power
  - Usage_kWh
  - Lagging_Current_Reactive.Power_kVarh
  - Leading_Current_Reactive_Power_kVarh
- **Set 2 (Y)**: Related to environmental factors and power factors
  - CO2.tCO2
  - Lagging_Current_Power_Factor
  - Leading_Current_Power_Factor

The CCA results revealed the following canonical correlations

```r
# Extract canonical correlations
rho <- cc_model$cor
rho
```

```
[1] 0.9887682 0.9365642 0.5055921
```

The first two canonical correlations show very strong relationships between the sets of variables, indicating that energy usage and reactive power are highly dependent on environmental factors and power factors.

The third canonical correlation of 0.5055921 indicates a moderate relationship. Although this correlation is weaker compared to the first two, it still represents a meaningful linear relationship, suggesting that some aspects of energy usage and reactive power are moderately related to environmental factors and power factors.

Squared Canonical Correlations

```r
# Squared canonical correlations
cc_model$cor[1:3 ]^2
```

```
[1] 0.9776625 0.8771526 0.2556234
```

The first two canonical correlations are very strong (97.8% and 87.7%), indicating that most of the variance in energy usage and reactive power can be explained by the environmental and power factors.

Test for the independence between Canonical Variate Pairs

```
Wilks' Lambda, using F-approximation (Rao's F):
                stat    approx df1     df2 p.value
1 to 3:  0.002042651 111234.76   9 85263.7       0
2 to 3:  0.091444750  40411.06   4 70070.0       0
3 to 3:  0.744376593  12031.57   1 35036.0       0
 Hotelling-Lawley Trace, using F-approximation:
              stat    approx df1    df2 p.value
1 to 3:  51.251268 199496.51   9 105098       0
2 to 3:   7.483585  65546.22   4 105104       0
3 to 3:   0.343406  12031.80   1 105110       0
 Pillai-Bartlett Trace, using F-approximation:
              stat    approx df1    df2 p.value
1 to 3:  2.1104384 27707.028   9 105108       0
2 to 3:  1.1327760 15942.197   4 105114       0
3 to 3:  0.2556234  9791.343   1 105120       0
 Roy's Largest Root, using F-approximation:
              stat   approx df1   df2 p.value
1 to 1:  0.9776625 511148.2   3 35036       0

 F statistic for Roy's Greatest Root is an upper bound.
```

The results of the Wilks' Lambda, Hotelling-Lawley Trace, Pillai-Bartlett Trace, and Roy's Largest Root tests indicate highly significant canonical correlations (p-values = 0). Given that all p-values are below the 1% significance level, we reject the null hypothesis of no significant correlations. This confirms strong relationships between energy usage and reactive power with environmental factors and power factors, highlighting their potential for optimizing energy efficiency in steel production.

Estimated Canonical Coefficients for Set 1

```
cc_model$xcoef
```

```
                                            [,1]         [,2]         [,3]
Usage_kWh                          -0.0052735175 2.256844e-03  0.010645322
Lagging_Current_Reactive.Power_kVarh  0.0001421719 6.903281e-05 -0.012507868
Leading_Current_Reactive_Power_kVarh  0.0005338996 5.633823e-03 -0.001556779
```

This shows the relationships between the variables in Set 1 and the first three canonical variates. Notably, **Leading_Current_Reactive_Power_kVarh** has a stronger positive relationship with the second canonical variate, while **Usage_kWh** shows a moderate positive relationship with the third canonical variate.

Estimated Canonical Coefficients for Set 2

```
```{r}
cc_model$ycoef
```
```

```
                                  [,1]          [,2]         [,3]
CO2.tCO2.                -0.0050579944  0.0021319794 -0.006019203
Lagging_Current_Power_Factor -0.0001390614  0.0005576595  0.008878646
Leading_Current_Power_Factor -0.0005881709 -0.0052743336  0.007048308
```

Leading_Current_Power_Factor has a notable negative relationship with the second canonical variate and a positive relationship with the third. **CO2.tCO2.** has small negative coefficients, indicating weak inverse relationships with the canonical variates.

Correlation between the Set 1 variables and the Canonical variables for Set 1

```
loadings$corr.X.xscores
```

```
                                      [,1]       [,2]         [,3]
Usage_kWh                       -0.9957577  0.09137659 -0.010814054
Lagging_Current_Reactive.Power_kVarh -0.8984971 -0.03575134 -0.437521281
Leading_Current_Reactive_Power_kVarh  0.4098986  0.91207959  0.009693057
```

Usage_kWh is highly negatively correlated with the first canonical variate, **Lagging_Current_Reactive.Power_kVarh** is strongly negatively correlated with the first and third canonical variates, and **Leading_Current_Reactive_Power_kVarh** is strongly positively correlated with the second canonical variate. This suggests that **Usage_kWh** and **Lagging_Current_Reactive.Power_kVarh** primarily influence the first canonical variate, while **Leading_Current_Reactive_Power_kVarh** has a strong influence on the second canonical variate.

Correlation between the Set 2 variables and the Canonical variables for Set 2

```{r}
loadings$corr.Y.yscores
```

	[,1]	[,2]	[,3]
CO2.tCO2.	-0.9963098	0.08326245	-0.02083429
Lagging_Current_Power_Factor	-0.3281896	0.76923546	0.54824119
Leading_Current_Power_Factor	-0.4374259	-0.89788860	0.04954439

CO2.tCO2. is highly negatively correlated with the first canonical variate, **Lagging_Current_Power_Factor** is strongly positively correlated with the second canonical variate, and **Leading_Current_Power_Factor** is strongly negatively correlated with the second canonical variate. This suggests that **CO2.tCO2.** primarily influences the first canonical variate, while **Lagging_Current_Power_Factor** and **Leading_Current_Power_Factor** have strong, but opposite, influences on the second canonical variate.

Correlation between the Set 1 variables and the Canonical variables for Set 2

```{r}
loadings$corr.X.yscores
```

	[,1]	[,2]	[,3]
Usage_kWh	-0.9845735	0.08558005	-0.005467501
Lagging_Current_Reactive.Power_kVarh	-0.8884053	-0.03348343	-0.221207319
Leading_Current_Reactive_Power_kVarh	0.4052947	0.85422113	0.004900733

Usage_kWh and **Lagging_Current_Reactive.Power_kVarh** significantly influence the first Y canonical variate, and **Leading_Current_Reactive_Power_kVarh** significantly influences the second Y canonical variate.

Correlation between the Set 2 variables and the Canonical variables for Set 1

```r
loadings$corr.Y.xscores
```

```
                                [,1]        [,2]        [,3]
CO2.tCO2.                 -0.9851194  0.07798063 -0.01053365
Lagging_Current_Power_Factor -0.3245034  0.72043842  0.27718643
Leading_Current_Power_Factor -0.4325128 -0.84093035  0.02504925
```

CO2.tCO2. primarily influences the first X canonical variate, while **Lagging_Current_Power_Factor** and **Leading_Current_Power_Factor** have strong, but opposite, influences on the second X canonical variate.

Conclusion and Recommendation

- **Strong Relationships:** CCA revealed significant links between energy usage/reactive power and environmental/power factors, with high canonical correlations (0.9888 and 0.9366).
- **Key Influencers: Usage_kWh** and **Lagging_Current_Reactive.Power_kVarh** are crucial for the first variate, while **Leading_Current_Reactive_Power_kVarh** impacts the second. **CO2.tCO2.** strongly affects the first variate, and power factors influence the second.
- **Statistical Confirmation:** Statistical tests (Wilks' Lambda, Hotelling-Lawley Trace, Pillai's Trace, Roy's Greatest Root) confirmed the significance of the canonical correlations, indicating strong dependencies between the two sets of variables.
- **Energy Optimization:** These findings can help optimize energy use, cut costs, and reduce environmental impact in the steel industry.
- **Recommendations:** Focus on key variables, manage CO2 and power factors, and continuously monitor energy use for better efficiency.

References

https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption

https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/

https://www.youtube.com/watch?v=Pul1BiXUda0

Appendices

Part of the data set.

	date <chr>	Usage_kWh <dbl>	Lagging_Current_Reactive.Power_kVarh <dbl>	Leading_Current_Reactive_Power_kVarh <dbl>	CO2.tCO2. <dbl>	Lagging_Current_Power_Factor <dbl>
1	01/01/2018 00:15	3.17	2.95	0	0	73.21
2	01/01/2018 00:30	4.00	4.46	0	0	66.77
3	01/01/2018 00:45	3.24	3.28	0	0	70.28
4	01/01/2018 01:00	3.31	3.56	0	0	68.09
5	01/01/2018 01:15	3.82	4.50	0	0	64.72

Leading_Current_Reactive_Power_kVarh <dbl>	CO2.tCO2. <dbl>	Lagging_Current_Power_Factor <dbl>	Leading_Current_Power_Factor <dbl>	NSM <int>	WeekStatus <chr>	Day_of_week <chr>	Load_Type <chr>
0	0	73.21	100	900	Weekday	Monday	Light_Load
0	0	66.77	100	1800	Weekday	Monday	Light_Load
0	0	70.28	100	2700	Weekday	Monday	Light_Load
0	0	68.09	100	3600	Weekday	Monday	Light_Load
0	0	64.72	100	4500	Weekday	Monday	Light_Load

12 of 11 columns

```
library(readr), library(dplyr) ,library(CCA), library(CCP)

data <- read.csv('../data/Steel_industry_data.csv')

head(data), colnames(data)

# Select relevant columns for Set 1 and Set 2

data1 <- data %>%

  select(Usage_kWh, Lagging_Current_Reactive.Power_kVarh, Leading_Current_Reactive_Power_kVarh)

data2 <- data %>%

  select(CO2.tCO2., Lagging_Current_Power_Factor, Leading_Current_Power_Factor)

# Ensure the data is numeric

data1 <- data1 %>%

  mutate(across(everything(), as.numeric))

data2 <- data2 %>%

  mutate(across(everything(), as.numeric))

# Check for any missing values

sum(is.na(data1))

sum(is.na(data2))

# Remove any rows with missing values
```

```
data1 <- na.omit(data1)
data2 <- na.omit(data2)
# Standardize the datasets
data1 <- scale(data1)
data2 <- scale(data2)
# Perform Canonical Correlation Analysis
cc_model <- cancor(data1, data2)
# Extract canonical correlations
rho <- cc_model$cor
# Number of observations and number of variables in each set
n <- nrow(data)
p <- ncol(data1)
q <- ncol(data2)
# Perform Wilks' lambda, Hotelling-Lawley trace, Pillai's trace, and Roy's greatest root tests
wilks_result <- p.asym(rho, n, p, q, tstat = "Wilks")
hotelling_result <- p.asym(rho, n, p, q, tstat = "Hotelling")
pillai_result <- p.asym(rho, n, p, q, tstat = "Pillai")
roy_result <- p.asym(rho, n, p, q, tstat = "Roy")
# Canonical correlations
cc_model$cor[1:3]
# Squared canonical correlations
cc_model$cor[1:3 ]^2
# Canonical coefficients for X and Y
cc_model$xcoef
cc_model$ycoef
# Compute loadings
loadings <- comput(data1, data2, cc_model)
# Correlations
loadings$corr.X.xscores
loadings$corr.Y.yscores
loadings$corr.X.yscores
loadings$corr.Y.xscores
```